

# Gradient-based kernel dimension reduction for supervised learning

Kenji Fukumizu\* and Chenlei Leng<sup>†</sup>

September 5, 2011

## Abstract

This paper proposes a novel kernel approach to linear dimension reduction for supervised learning. The purpose of the dimension reduction is to find directions in the input space to explain the output as effectively as possible. The proposed method uses an estimator for the gradient of regression function, based on the covariance operators on reproducing kernel Hilbert spaces. In comparison with other existing methods, the proposed one has wide applicability without strong assumptions on the distributions or the type of variables, and uses computationally simple eigendecomposition. Experimental results show that the proposed method successfully finds the effective directions with efficient computation.

## 1 Introduction

Dimension reduction is involved in most of modern data analysis, in which high dimensional data must be often handled. The purpose of dimension reduction is multifold: preprocessing for another data analysis, aiming at less expensive computation in later processing, or construction of readable low dimensional expressions. There are two categories of dimension reduction: unsupervised methods such as PCA, and supervised methods such as Fisher discriminant analysis (FDA). This paper focuses on dimension reduction for supervised learning.

---

\*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562 Japan

<sup>†</sup>Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore, 117546

Let  $(X, Y)$  be a random vector such that  $X$  takes values in  $\mathbb{R}^m$ . The domain of  $Y$  can be arbitrary, either continuous, discrete, or structured. Supervised learning concerns how  $Y$  is explained by  $X$ . The purpose of dimension reduction in this setting is to find such features of  $X$  that explain  $Y$  as effectively as possible. This paper focuses linear dimension reduction, in which linear combinations of the components of  $X$  are used to make effective features. Although there are many methods for extracting nonlinear features including kernel methods, this paper confines its attentions on linear features for the following reasons: (i) nonlinear feature extraction such as kernel method depends strongly on the choice of the nonlinearity (see Sec. 3.2, Wine data, for example). Linear methods are more stable. (ii) we can apply some nonlinear transform  $\phi(X)$  of  $X$  so that linear combinations of  $\phi(X)$  give effective features of  $X$ , once a linear dimension reduction method is established.

Beyond the classical approaches such as FDA and CCA, the modern approach to this linear dimension reduction is based on the formulation by conditional independence. More precisely, we assume

$$p(Y|X) = \tilde{p}(Y|B^T X) \quad \text{or equivalently} \quad Y \perp\!\!\!\perp X | B^T X \quad (1)$$

for the distribution, where  $B$  is a projection matrix ( $B^T B = I_d$ ) onto a  $d$ -dimensional subspace ( $d < m$ ) in  $\mathbb{R}^m$ , and wish to estimate  $B$ . The subspace spanned by the column vectors of  $B$  is called the *effective direction for regression*, or *EDR space* [14]. We consider methods of estimating  $B$  without specific parametric models for  $p(y|x)$ , unlike the model-based approach such as [15]

The first method that aims at finding the EDR space is the *sliced inverse regression* (SIR, [13]), which employs the fact that the inverse regression  $E[X|Y]$  lies in the EDR space under some assumptions. Many methods have been proposed in this vein of inverse regression ([3, 12]), which use some statistic in each slice of  $Y$ . While many inverse regression methods are computationally simple, they often need some strong assumptions on the distribution of  $X$  such as elliptic symmetry, and slice-based methods are not effective for classification, where the number of slices is at most that of classes. Another interesting approach is the minimum average variance estimation (MAVE [21]), in which the conditional variance of the regression in the direction of  $B^T X$ ,  $E[(Y - E[Y|B^T X])^2 | B^T X]$ , is minimized with the conditional variance estimated by the local linear kernel smoothing method. The kernel smoothing method requires, however, careful choice of bandwidth parameter, and it is usually difficult to apply if the dimensionality is very high.

The most relevant to this paper is the methods that use the gradient of regressor  $\varphi(x) = E[Y|X = x]$  [16, 11]. As explained in Sec. 2.1, under Eq. (1) the gradient of  $\varphi(x)$  is contained in the EDR space. One can estimate the space by nonparametric estimation of the gradient. There are some limitations in this method, however: the nonparametric estimation of the gradient in high-dimensional spaces is challenging, and the gradient is not estimable if some symmetry holds in the system.

A kernel method for dimension reduction has been proposed to overcome various limitations of existing methods. The kernel dimension reduction (KDR, [7, 8, 20]) uses the kernel method to characterize the conditional independence relation in Eq. (1). While KDR is a general method applicable to a wide class of problems without requiring any strong assumptions on the distributions or types of  $X$  or  $Y$ , the optimization needed for the estimation is computationally a problem: the objective function is non-convex, and the gradient descent method demands many inversions of Gram matrices, which prohibits applications to very high-dimensional or large data.

We propose a novel kernel method for dimension reduction using the gradient-based approach, but unlike the existing ones [16, 11], the gradient is estimated by the covariance operators with positive definite kernels, which is based on the recent development in the kernel method [8, 17]. It solves the problems of existing methods: by virtue of the kernel method the response  $Y$  can be of arbitrary type, and the kernel estimation of the gradient is stable without careful decrease of bandwidth. It solves also the problem of KDR: the estimator by an eigenproblem needs no numerical optimization. The method is thus applicable to large and high-dimensional data, as we demonstrate experimentally.

## 2 Gradient-based kernel dimension reduction

In this paper, the range of an operator  $A$  is denoted by  $\mathcal{R}(A)$ .

### 2.1 Gradient of a regression function and dimension reduction

We first review the basic idea of the gradient-based method for dimension reduction in supervised learning, which has been used in [16, 11]. Suppose  $Y$  is a real-valued random variable such that the regression function  $E[Y|X =$

$x]$  is differentiable w.r.t.  $x$ . If the assumption Eq. (1) holds, we have

$$\frac{\partial}{\partial x} E[Y|X = x] = \frac{\partial}{\partial x} \int yp(y|x)dy = \int y \frac{\partial \tilde{p}(y|B^T x)}{\partial x} dy = B \int y \frac{\partial \tilde{p}(y|z)}{\partial z} \Big|_{z=B^T x} dy,$$

which implies that the gradient  $\frac{\partial}{\partial x} E[Y|X = x]$  at any  $x$  is contained in the EDR space. Based on this fact, the average derivative estimates (ADE, [16]) has been proposed to use the average of the gradients for estimating  $B$ . In the more recent method [11], assuming that  $Y$  is one-dimensional continuous variable, a standard local linear least squares with a smoothing kernel (not necessarily positive definite kernel) [4] is used for estimating the gradient, and the dimensionality of the projection is iteratively reduced to the desired one. Since the gradient estimation for high-dimensional data is difficult in general, the iterative reduction is expected to give a more accurate estimation. We call the method in [11] iterative average derivative estimates (IADE).

## 2.2 Kernel method for conditional expectation

It has been recently revealed that the apparatus of positive definite kernels or reproducing kernel Hilbert space (RKHS) can be applied to estimate the regression function or conditional expectation with covariance operators on RKHS [7, 8, 17], which we briefly review below. For a set  $\Omega$ , a ( $\mathbb{R}$ -valued) positive definite kernel  $k$  on  $\Omega$  is a symmetric kernel  $k : \Omega \times \Omega \rightarrow \mathbb{R}$  such that  $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$  for any  $x_1, \dots, x_n$  in  $\Omega$  and  $c_1, \dots, c_n \in \mathbb{R}$ . It is known that a positive definite kernel on  $\Omega$  uniquely defines a Hilbert space  $\mathcal{H}$  consisting of functions on  $\Omega$  such that (i)  $k(\cdot, x)$  is in  $\mathcal{H}$ , (ii) the linear hull of  $\{k(\cdot, x) \mid x \in \Omega\}$  is dense in  $\mathcal{H}$ , and (iii) for any  $x \in \Omega$  and  $f \in \mathcal{H}$ ,  $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$  (reproducing property), where  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  is the inner product of  $\mathcal{H}$ . The Hilbert space  $\mathcal{H}$  is called the *reproducing kernel Hilbert space* (RKHS) associated with  $k$ .

Let  $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu_{\mathcal{X}})$  and  $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}}, \mu_{\mathcal{Y}})$  be measure spaces, and  $(X, Y)$  be a random variable on  $\mathcal{X} \times \mathcal{Y}$  with probability  $P$ . We assume that the probability density function (p.d.f.)  $p(x, y)$  and the conditional p.d.f.  $p(y|x)$  always exist. Also, we always assume that a positive definite kernel is measurable and bounded: the boundedness means  $\sup_{x \in \Omega} k(x, x) < \infty$ .

Let  $k_{\mathcal{X}}$  and  $k_{\mathcal{Y}}$  be positive definite kernels on  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively, with respective RKHS  $\mathcal{H}_{\mathcal{X}}$  and  $\mathcal{H}_{\mathcal{Y}}$ . The (uncentered) *covariance operator*  $C_{YX} : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$  is defined by the equation

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = E[f(X)g(Y)] = E[\langle f, \Phi_{\mathcal{X}}(X) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle \Phi_{\mathcal{Y}}(Y), g \rangle_{\mathcal{H}_{\mathcal{Y}}}] \quad (2)$$

for all  $f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}$ , where  $\Phi_{\mathcal{X}}(x) = k_{\mathcal{X}}(\cdot, x)$  and  $\Phi_{\mathcal{Y}}(y) = k_{\mathcal{Y}}(\cdot, y)$ . Similarly,  $C_{XX}$  denotes the operator on  $\mathcal{H}_{\mathcal{X}}$  that satisfies  $\langle f_2, C_{XX}f_1 \rangle = E[f_2(X)f_1(X)]$  for any  $f_1, f_2 \in \mathcal{H}_{\mathcal{X}}$ . These definitions are straightforward extensions of the ordinary covariance matrices, if we consider the covariance of the random vectors  $\Phi_{\mathcal{X}}(X)$  and  $\Phi_{\mathcal{Y}}(Y)$  on RKHS.

By setting  $g = k_{\mathcal{Y}}(\cdot, y)$  in Eq. (2), the reproducing property derives

$$(C_{YX}f)(y) = \int k_{\mathcal{Y}}(y, \tilde{y})f(\tilde{x})dP(\tilde{x}, \tilde{y}), \quad (C_{XX}f)(x) = \int k_{\mathcal{X}}(x, \tilde{x})f(\tilde{x})dP_X(\tilde{x}),$$

which shows the explicit expressions of  $C_{YX}$  and  $C_{XX}$  as integral operators.

An advantage of the kernel method is that estimation with finite data is straightforward. Given i.i.d. sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  with law  $P$ , the covariance operator is estimated by

$$\hat{C}_{YX}^{(n)}f = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{Y}}(\cdot, Y_i) \langle k_{\mathcal{X}}(\cdot, X_i), f \rangle_{\mathcal{H}_{\mathcal{X}}} = \frac{1}{n} \sum_{i=1}^n f(X_i) k_{\mathcal{Y}}(\cdot, Y_i). \quad (3)$$

The estimator  $\hat{C}_{XX}^{(n)}$  is given similarly. It is known that these estimators are  $\sqrt{n}$ -consistent in Hilbert-Schmidt norm [10].

The fundamental result in discussing conditional probabilities with kernels is the following fact.

**Theorem 1** ([7]). *If  $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$  holds for  $g \in \mathcal{H}_{\mathcal{Y}}$ , then*

$$C_{XX}E[g(Y)|X = \cdot] = C_{XY}g.$$

If  $C_{XX}$  is injective<sup>1</sup>, the above relation can be expressed as

$$E[g(Y)|X = \cdot] = C_{XX}^{-1}C_{XY}g. \quad (4)$$

The assumption  $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$  may not hold in general; we can easily make counterexamples with Gaussian kernel and Gaussian distributions. We can nonetheless obtain an empirical estimator based on Eq. (4), namely,

$$(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{C}_{XY}^{(n)}g,$$

where  $\varepsilon_n$  is a regularization coefficient in Thikonov-type regularization. As we discuss in Appendix, we can in fact prove rigorously that this estimator converges to  $E[g(Y)|X = \cdot]$ .

---

<sup>1</sup>Noting  $\langle C_{XX}f, f \rangle = E[f(X)^2]$ , it is easy to see that  $C_{XX}$  is injective, if  $k_{\mathcal{X}}$  is a continuous kernel on a topological space  $\mathcal{X}$ , and  $P_X$  is a Borel probability measure such that  $P(U) > 0$  for any open set  $U$  in  $\mathcal{X}$ .

To apply the above kernel expressions to the method discussed in Sec. 2.1, we need a way of taking the derivative of a function. It is known (*e.g.*, [19] Sec. 4.3) that if a positive definite kernel  $k(x, y)$  on an open set in Euclidean space is continuously differentiable with respect to  $x$  and  $y$ , every  $f$  in the corresponding RKHS is continuously differentiable. If further  $\frac{\partial}{\partial x}k(\cdot, x) \in \mathcal{H}_{\mathcal{X}}$ , we have

$$\frac{\partial f}{\partial x} = \left\langle f, \frac{\partial}{\partial x}k(\cdot, x) \right\rangle_{\mathcal{H}_{\mathcal{X}}}. \quad (5)$$

Namely, the derivative of any function in that RKHS can be computed in the form of the inner product. This property combined with the above kernel estimator of  $E[g(Y)|X = x]$  provides a method for dimension reduction.

## 2.3 Gradient-based kernel method for dimension reduction

### 2.3.1 Algorithm

Assume that  $\mathcal{X} = \mathbb{R}^m$ ,  $C_{XX}$  is injective,  $k_{\mathcal{X}}(x, \tilde{x})$  is continuously differentiable,  $E[g(Y)|X = x] \in \mathcal{H}_{\mathcal{X}}$  for any  $g \in \mathcal{H}_{\mathcal{Y}}$ , and  $\frac{\partial}{\partial x}k_{\mathcal{X}}(\cdot, x) \in \mathcal{R}(C_{XX})$ . It follows from Eqs. (4) and (5) that

$$\frac{\partial}{\partial x}E[g(Y)|X = x] = \left\langle C_{XX}^{-1}C_{XY}g, \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x} \right\rangle = \left\langle g, C_{YX}C_{XX}^{-1} \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x} \right\rangle. \quad (6)$$

Define  $\Psi : \mathbb{R}^m \rightarrow \mathcal{H}_{\mathcal{Y}}$ ,  $x \mapsto E[k_{\mathcal{Y}}(\cdot, Y)|X = x]$ . By plugging  $g = k(\cdot, y)$  into Eq. (6), we see

$$\frac{\partial \Psi(x)}{\partial x} = C_{YX}C_{XX}^{-1} \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x}.$$

On the other hand, from  $\Psi(x) = \int k_{\mathcal{Y}}(\cdot, y)p(y|x)d\mu_y(y)$ , the same argument as in Sec. 2.1 shows that  $\frac{\partial \Psi(x)}{\partial x} = \Xi(x)B$  with an operator  $\Xi(x)$  from  $\mathbb{R}^m$  to  $\mathcal{H}_{\mathcal{Y}}$ , where we use a slight abuse of notation by identifying the operator  $\Xi(x)$  with a matrix. Taking the inner product in  $\mathcal{H}_{\mathcal{Y}}$ , we have

$$B^T \langle \Xi(x), \Xi(x) \rangle_{\mathcal{H}_{\mathcal{Y}}} B = \left\langle \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x}, C_{XX}^{-1}C_{XY}C_{YX}C_{XX}^{-1} \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x} \right\rangle =: M(x),$$

which shows that the eigenvectors for non-zero eigenvalues of the  $m \times m$  symmetric matrix  $M(x)$  are contained in the EDR space. This fact is the basis of the proposed method. Note that, in comparison with the conventional gradient-based method described in Sec. 2.1, this method is interpreted as considering simultaneously various regression functions  $E[k_{\mathcal{Y}}(\tilde{y}, Y)|X = x]$  given by all  $\tilde{y} \in \mathcal{Y}$ .

Given i.i.d. sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from the true distribution, based on the empirical covariance operators Eq. (3) and regularized inversions, the matrix  $M(x)$  is estimated by

$$\begin{aligned}\widehat{M}_n(x) &= \left\langle \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x}, (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{C}_{XY}^{(n)} \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x} \right\rangle \\ &= \nabla \mathbf{k}_X(x)^T (G_X + n\varepsilon_n I)^{-1} G_Y (G_X + n\varepsilon_n I)^{-1} \nabla \mathbf{k}_X(x),\end{aligned}\quad (7)$$

where  $G_X$  and  $G_Y$  are Gram matrices  $(k_{\mathcal{X}}(X_i, X_j))$  and  $(k_{\mathcal{Y}}(Y_i, Y_j))$ , respectively, and  $\nabla \mathbf{k}_X(x) = (\frac{\partial k_{\mathcal{X}}(X_1, x)}{\partial x}, \dots, \frac{\partial k_{\mathcal{X}}(X_n, x)}{\partial x})^T \in \mathbb{R}^n$ .

As the eigenvectors of  $M(x)$  are contained in the EDR space for any  $x$ , we propose to use the average of  $M(X_i)$  over all the data points  $X_i$ , and define

$$\tilde{M}_n = \frac{1}{n} \sum_{i=1}^n \widehat{M}_n(X_i) = \frac{1}{n} \sum_{i=1}^n \nabla \mathbf{k}_X(X_i)^T (G_X + n\varepsilon_n I)^{-1} G_Y (G_X + n\varepsilon_n I)^{-1} \nabla \mathbf{k}_X(X_i).$$

In the case of Gaussian kernel, for example,  $\nabla \mathbf{k}_X(X_i)$  is given by  $(X_i - X_j) \exp(-\frac{1}{2\sigma^2} \|X_i - X_j\|^2)$ , which is the Hadamard product between the Gram matrix  $G_X$  and  $(X_i - X_j)_{ij=1}^n$ .

The projection matrix  $B$  in Eq. (1) is then estimated by the top  $d$  eigenvectors of the  $m \times m$  symmetric matrix  $\tilde{M}_n$ . We call this method *gradient-based kernel dimension reduction* (gKDR).

### 2.3.2 Discussions and extensions

The proposed gKDR applies to a wide class of problems. In contrast to many existing methods, the gKDR can handle any type of data for  $Y$  including multivariate or structured variables, and make no strong assumptions on the distribution of  $X$ . The gKDR method can be applied to classification and continuous output exactly in the same manner.

The previous gradient-based methods ADE and IADE have an obvious weakness. Suppose  $Y$  is one-dimensional and  $Y = \varphi(B^T X) + Z$ , where  $Z$  is a zero-mean noise. If  $E[\varphi'(B^T X)] = 0$ , the subspace spanned by  $B$  cannot be estimated. This condition holds if  $\varphi$  and the distribution of  $X$  satisfy some symmetry. These methods in general find only a subspace of the EDR space. In contrast, the gKDR approach incorporates various functions  $k_{\mathcal{Y}}(\tilde{y}, \cdot)$  for  $\varphi$ , as discussed in Sec. 2.3.1, and thus this weakness may be avoided.

As in all kernel methods, the results of gKDR depend on the choice of kernels, though the linear features are less sensitive to the choice than nonlinear features. We use the cross-validation (CV) for choosing kernels and parameters, combined with some regression or classification method. In this paper, the k-nearest neighbor (kNN) regression / classification is used

in CV for its simplicity: for each candidate of a kernel or parameter, we compute the CV error by the kNN method with the input data projected on the subspace given by gKDR, and choose the one that gives the least error.

The time complexity of the matrix inversions and the eigendecomposition required for gKDR are  $O(n^3)$ , which is prohibitive for large data sets. We can apply, however, low-rank approximation of Gram matrices, such as incomplete Cholesky decomposition [5], which is a standard method for reducing time complexity in kernel methods. The space complexity may be also a problem of gKDR, since  $(\nabla \mathbf{k}_X(X_i))_{i=1}^n$  has  $n^2 \times m$  dimension. In the case of Gaussian kernel, we have a way of reducing the necessary memory by low rank approximation of the Gram matrices. Note that  $\frac{\partial}{\partial x^a} k_X(X_j, x)|_{x=X_i}$  for Gaussian kernel is given by  $\frac{1}{\sigma^2}(X_j^a - X_i^a) \exp(-\|X_j - X_i\|^2/(2\sigma^2))$ . Let  $G_X \approx RR^T$  and  $G_Y \approx HH^T$  be the low rank approximation with  $r_x = \text{rk}R, r_y = \text{rk}H$  ( $r_x, r_y < n, m$ ). With the notation  $F := (G_X + n\varepsilon_n I_n)^{-1}H$  and  $\Theta_i^{as} = \frac{1}{\sigma^2}X_i^a R_{is}$ , we have

$$\tilde{M}_{n,ab} = \sum_{i=1}^n \sum_{t=1}^{r_y} \Gamma_{ia}^t \Gamma_{ib}^t \quad (1 \leq a, b \leq m),$$

$$\Gamma_{ia}^t = \sum_{j=1}^n \sum_{s=1}^{r_x} \frac{1}{\sigma^2} (X_j^a - X_i^a) R_{js} R_{is} F_{jt} = \sum_{s=1}^{r_x} R_{is} \left( \sum_{j=1}^n \Theta_j^{as} F_{jt} \right) - \sum_{s=1}^{r_x} \Theta_i^{as} \left( \sum_{j=1}^n R_{js} F_{jt} \right).$$

With this method, the complexity is  $O(nmr)$  in space and  $O(nm^2r)$  in time ( $r = \max\{r_x, r_y\}$ ), which is much more efficient in memory than straightforward implementation.

We introduce two variants of gKDR. First, as discussed in [11], accurate nonparametric estimation for the derivative of regression function with high-dimensional  $X$  is not easy in general. We propose a method for decreasing the dimensionality iteratively in a similar manner to IADE. Using gKDR, we first find a projection matrix  $B_1$  of a larger dimension  $d_1$  than the target dimensionality  $d$ , project data  $X_i$  onto the subspace as  $Z_i^{(1)} = B_1^T X_i$ , and find the projection matrix  $B_2$  ( $d_1 \times d_2$  matrix) for  $Z_i^{(1)}$  onto a  $d_2$  ( $d_2 < d_1$ ) dimensional subspace. After repeating this process to the dimensionality  $d$ , the final result is given by  $\hat{B} = B_\ell \cdots B_2 B_1$ . In this way, we can expect the later projector is more accurate by the low dimensionality of the data  $Z_i^{(s)}$ . We call this method gKDR-i.

Second, in classification problems, where the  $L$  classes are encoded as  $L$  different points, the Gram matrix  $G_Y$  is of rank  $L$  at most. We can have at most  $L$  dimensional subspace by the gKDR method (see Eq. (7)), which



is a strong limitation of gKDR, especially for binary classification. Note that this problem is shared by many linear dimension reduction methods including CCA and slice-based methods. To solve this problem, we propose to use the variation of  $\widehat{M}_n(x)$  over all points  $x = X_i$  instead of the average  $\tilde{M}_n$ . We compute the projection matrix  $\widehat{B}_i$  from  $\widehat{M}_n(X_i)$  at each  $i$ , take the average of projectors  $\widehat{P} = \frac{1}{n} \sum_{i=1}^n \widehat{B}_i \widehat{B}_i^T$ , and give the estimator  $B$  by the top eigenvectors of  $\widehat{P}$ . In practice, eigendecomposition of  $\widehat{M}(X_i)$  for all  $i$  may not be feasible. In that case, by partitioning  $\{1, \dots, n\}$  into  $T_1, \dots, T_\ell$ , the projection matrices  $\widehat{B}_{[a]}$  given by the eigenvectors of  $\widehat{M}_{[a]} = \sum_{i \in T_a} \widehat{M}(X_i)$  can be used to define  $\widehat{P} = \frac{1}{\ell} \sum_{a=1}^{\ell} \widehat{B}_{[a]} \widehat{B}_{[a]}^T$ . We call this method gKDR-v.

### 2.3.3 Theoretical analysis of gKDR

Under some conditions, we can obtain the consistency and its rate for  $\widehat{M}_n(x)$  and  $\tilde{M}_n$ . We assume all the RKHS are separable, and  $\|M\|_F$  denotes Frobenius norm of a matrix  $M$ .

**Theorem 2.** *Assume that  $\frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x^a} \in \mathcal{R}(C_{XX}^{\beta+1})$  ( $a = 1, \dots, m$ ) for some  $\beta \geq 0$  and  $E[k_Y(y, Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$  for every  $y \in \mathcal{Y}$ . Then, for  $\varepsilon_n = n^{-\max\{\frac{1}{3}, \frac{1}{2\beta+2}\}}$ , we have*

$$\widehat{M}_n(x) - M(x) = O_p\left(n^{-\min\{\frac{1}{3}, \frac{2\beta+1}{4\beta+4}\}}\right)$$

for every  $x \in \mathcal{X}$  as  $n \rightarrow \infty$ . If further  $E[\|M(X)\|_F^2] < \infty$  and  $\frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x^a} = C_{XX}^{\beta+1} h_x^a$  with  $E\|h_X^a\|_{\mathcal{H}_{\mathcal{X}}} < \infty$ , then  $\tilde{M}_n \rightarrow E[M(X)]$  in the same order as above.

The proof is given in Appendix. Note that, assuming that the eigenvalues of  $M(x)$  or  $E[M(X)]$  are all distinct, the convergence of matrices implies the convergence of the eigenvectors, thus the estimator of gKDR is consistent to the subspace given by the top eigenvectors of  $E[M(X)]$ .

## 3 Experimental results

We always use the Gaussian kernel  $k(x, \tilde{x}) = \exp(-\frac{1}{2\sigma^2}\|x - \tilde{x}\|^2)$  in the kernel method below.

### 3.1 Synthesized data

First we use two types of synthesized data, which have been used in [11], to verify the basic performance of gKDR and the two variants. The data are

	gKDR	gKDR-i	gKDR-v	gKDR+KDR	IADE [11]
(A) $n = 100$	0.2114 (0.0636)	0.1905 (0.0495)	0.2101 (0.0704)	0.0883 (0.1473)	0.0903
(A) $n = 200$	0.1393 (0.0362)	0.1217 (0.0352)	0.1356 (0.0351)	0.0501 (0.0964)	0.0537
(B) $n = 100$	0.1500 (0.0363)	0.1358 (0.0347)	0.1630 (0.0398)	0.1076 (0.0967)	0.182
(B) $n = 200$	0.0755 (0.0157)	0.0750 (0.0153)	0.0802 (0.0160)	0.0506 (0.0729)	0.0472

Table 1: Synthesized data. Mean and standard error (in brackets) over 100 samples. The mean errors of IADE are taken from [11].

generated by

$$(A) : \quad Y = Z \sin(\sqrt{5}Z) + W, \quad Z = \frac{1}{\sqrt{5}}(1, 2, 0, \dots, 0)^T X,$$

$$(B) : \quad Y = (Z_1^3 + Z_2)(Z_1 - Z_2^3) + W, \\ Z_1 = \frac{1}{\sqrt{2}}(1, 1, 0, \dots, 0)^T X, \quad Z_2 = \frac{1}{\sqrt{2}}(1, -1, 0, \dots, 0)^T X,$$

where 10-dimensional  $X$  is generated by the uniform distribution on  $[-1, 1]^{10}$  and  $W$  is independent Gaussian noise with zero mean and variance  $10^{-2}$ . The sample size is  $n = 100$  and  $200$ . The discrepancy between the estimator  $B$  and the true projector  $B_0$  is measured by  $\|B_0 B_0^T (I_m - B B^T)\|_F / d$ , where  $\|\cdot\|_F$  is the Frobenius norm. For choosing the parameter  $\sigma$  in Gaussian kernel, CV with kNN ( $k = 5$ ) is used with 8 points given by  $c\sigma_{med}$  ( $0.5 \leq c \leq 10$ ), where  $\sigma_{med}$  is the median of pairwise distances of data [9] (the same strategy is used for CV in all the experiments below). The regularization parameter is fixed as  $\varepsilon_n = 10^{-7}$ .

We compare the results only with IADE, since [11] reports that the results of IADE are much better than those of SIR and pHd. From Table 1, we see that gKDR, gKDR-i (5 iterations), and gKDR-v show comparable results for data (B), while IADE works better for data (A). For data (B), when the sample size is 100, the proposed gKDR methods show much better results than IADE. gKDR and gKDR-v show similar errors, and gKDR-i improves them in all the four cases. We also use the results of gKDR as the initial state for KDR, which requires non-convex optimization with gradient method. As we can see from the table, KDR improves the accuracy significantly, showing results better than or comparable to IADE. The optimization in KDR, however, sometimes fails to find a good solution, which causes the large variance in the experiments.

	Dim.	Train	Test
heart-disease	13	149	148
ionosphere	34	151	200
breast-cancer	30	200	369

Table 2: Summary of data sets: dimensionality of  $X$  and the number of data

### 3.2 Real world data

We first use *Wine data*, which is available at the UCI machine learning repository [6], to demonstrate low dimensional visualization. In this data set,  $X$  is a 13 dimensional continuous variable, and  $Y$  is the class label representing three classes of wine, which is encoded as  $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ . The sample size is 173. Two dimensional projections are estimated by gKDR and KDR. For gKDR, the parameter  $\sigma$  in Gaussian kernel is chosen by CV with kNN ( $k = 5$ ). As in Figure 1, the results by the KDR and gKDR look similar, while each of the classes by KDR is more condensed. With Intel (R) Core (TM) i7 960, 3.20GHz, the computational time required for one parameter set was 0.14 sec by gKDR and 4.80 sec by KDR with 50 iterations of line search: gKDR is 30 times faster than KDR for this data set. As comparison, we show also the results by kernel CCA (KCCA) [1, 2]. Since the nonlinear mapping in KCCA easily separates the three classes with small  $\sigma$ , cross-validation is unstable and inapplicable. The results given by the three values of  $\sigma$  are very different for KCCA.

One way of evaluating dimension reduction methods in supervised learning is to consider the classification or regression accuracy after projecting data onto the estimated subspaces. We next use three data sets for binary classification, *heart-disease*, *ionosphere*, and *breast-cancer-Wisconsin*, from UCI repository (see Table 2), and compare the classification errors with gKDR-v and KDR.

The classification rates with kNN classifiers ( $k = 7$ ) for projected data are shown in Fig. 2. We can see that the classification ability of estimated subspaces by gKDR-v is competitive to those given by KDR: slightly worse in Ionosphere, and slightly better in Breast-cancer-Wisconsin. The computation of gKDR-v for these data sets can be hundreds or thousands times faster than that of KDR. For each parameter set, the computational time of gKDR vs KDR was, in *Heart-disease* 0.044 sec / 622 sec ( $d = 20$ ), in *Ionosphere* 0.103 sec / 84.77 sec ( $d = 20$ ), and in *Breast-cancer-Wisconsin* 0.116 sec / 615 sec ( $d = 11$ ).

The next two data sets are larger in the sample size and dimensionality, for which the optimization of KDR is difficult to apply. The first one is 2007

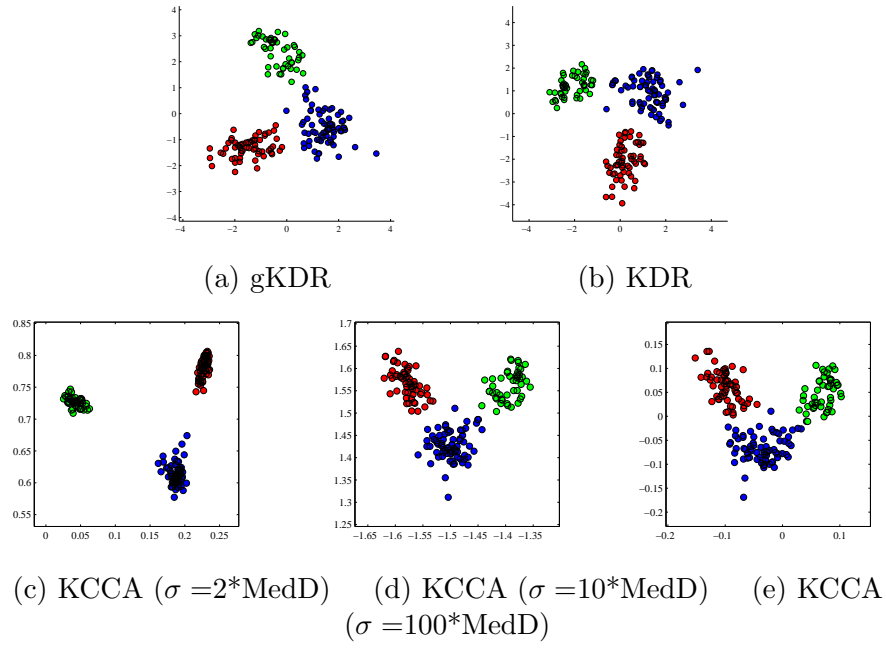


Figure 1: Two dimensional plots of Wine data by gKDR, KDR, and Kernel CCA. MedD means the median of pairwise distances among  $X_i$  [9].

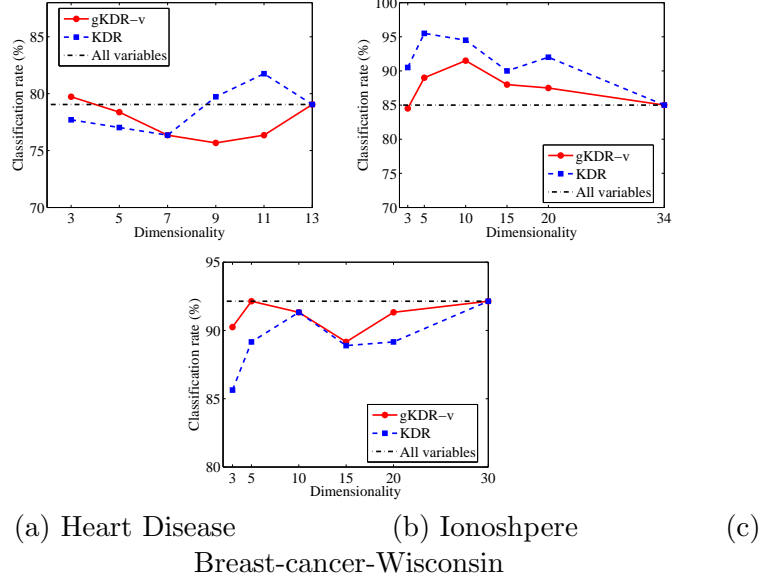


Figure 2: Classification accuracy with gKDR-v and KDR for binary classification problems

images of USPS handwritten digit data set used in [18], where 256 gray scale pixels are provided as  $X$  for each image. First we make a three dimensional plot for the subset of 500 images with classes “1” through “5”, as in the similar way to [20]. The result is shown in Fig. 3. We can see, although this is a linear projection, the subspace found by gKDR separates the five classes reasonably well.

We evaluate the classification errors by the simple kNN classifier ( $k = 5$ ) with the data projected onto estimated subspaces, using 1000 images for training and the rest for testing. We compare gKDR with CCA as a baseline. Table 3 shows that the subspaces found by gKDR (-i,-v) have much better classification ability than those given by CCA. As in the previous cases, gKDR and gKDR-v show similar errors, and gKDR-i (5 iterations) improves them slightly.

The second large data set is ISOLET, taken from UCI repository [6]. The data set provides 617 dimensional continuous features of speech signals for each of 26 alphabets. In addition to 6238 training data, 1559 test data are separately provided. We evaluate the classification errors with the kNN classifier ( $k = 5$ ) to see the effectiveness of the estimated subspaces. Table 4 shows the error rates of classification for the test data after dimension

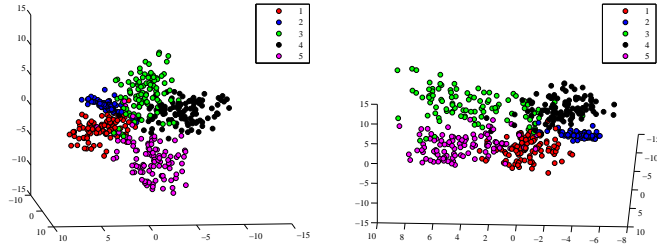


Figure 3: Three dimensional plots of USPS data (5 classes) from two different angles.

Dim.	3	5	7	9	15	20	25
gKDR	56.82	27.96	19.00	16.66	—	—	—
gKDR-i	39.81	26.17	18.62	15.06	—	—	—
gKDR-v	47.78	25.89	18.62	15.92	12.43	11.73	12.67
CCA	51.05	32.62	23.96	24.49	—	—	—

Table 3: USPS2007: classification errors for test data (percentage)

reduction. To save computational time, we did not use gKDR-i. From the information on the data at the UCI repository, the best performance with neural networks and C4.5 with ECOC are 3.27% and 6.61%, respectively. In comparison with these results, we can see the simple kNN classification shows competitive performance on the low dimensional subspaces found by gKDR and gKDR-v.

## 4 Concluding remarks

We have proposed a method for gradient-based kernel dimension reduction and its two variants, which provide general approach for dimension reduction in supervised learning; they have wide applicability with little restriction on the distribution or type of the variables, and the computation is done with simple linear algebra.

Dim.	5	10	15	20	25	30	35	40	45	50
gKDR	30.21	13.53	7.70	4.55	4.23	—	—	—	—	—
gKDR-v	29.44	13.15	8.28	4.55	3.91	4.81	5.26	5.26	5.77	5.58
CCA	22.77	15.78	8.72	6.74	7.18	—	—	—	—	—

Table 4: ISOLET: classification errors for test data (percentage)

As discussed in Sec. 2.3.2, gKDR may solve the problem of the existing gradient methods that they do not work if the regression function has the degenerate average derivative. It is then interesting to make a theoretical question whether gKDR can find the true EDR space. This is within our future works.

This paper focuses only on the supervised setting, but it may be possible to extend the proposed method to the unsupervised cases in a similar way employed in [20]. Extension to nonlinear feature extraction is also important in some practical problems. As we discuss in Introduction, applying a nonlinear transform will give a straightforward extension. Another interesting question is how we can “kernelize” gKDR to replace the linear features to nonlinear ones. This is not as straightforward as many other kernel methods, since the differentiation with respect to the feature map is involved. This is also within our interesting future directions.

## A Consistency of the kernel estimator for the regression function

We discuss the consistency of the estimator  $(\hat{C}_{XX} + \varepsilon_n I)^{-1} \hat{C}_{XY}^{(n)} g$  for  $E[g(Y)|X = \cdot]$ . While this consistency has been already proved in some literature such as [25, 26, 23, 24] in various contexts, we show the proof in our terminology for completeness.

**Theorem 3.** *Let  $g \in \mathcal{H}_Y$  and assume that  $E[g(Y)|X = \cdot] \in \mathcal{R}(C_{XX}^\nu)$  for  $\nu \geq 0$ , where  $\mathcal{R}(C_{XX}^0)$  for  $\nu = 0$  is interpreted as  $\mathcal{H}_X$ . If  $\varepsilon_n \rightarrow 0$  ( $n \rightarrow \infty$ ), then*

$$\|(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{C}_{XY}^{(n)} g - E[g(Y)|X = \cdot]\|_{\mathcal{H}_X}$$

*is of the order*

$$\begin{cases} O_p(\varepsilon_n^{-1} n^{-1/2}) + O(\varepsilon_n^\nu), & \text{for } 0 \leq \nu < 1, \\ O_p(\varepsilon_n^{-1} n^{-1/2}) + O(\varepsilon_n), & \text{for } \nu \geq 1. \end{cases}$$

*Consequently, if  $\varepsilon_n = n^{-\max\{\frac{1}{4}, \frac{1}{2\nu+2}\}}$ , then the estimator is consistent of the order  $O(n^{-\min\{\frac{1}{4}, \frac{\nu}{2\nu+2}\}})$ .*

*Proof.* Take  $\eta \in \mathcal{H}_X$  such that  $E[g(Y)|X = \cdot] = C_{XX}^\nu \eta$ . From Theorem 1, we have  $C_{XY} g = C_{XX} E[g(Y)|X = \cdot] = C_{XX}^{\nu+1} \eta$ .

First, we show

$$\|(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{C}_{XY}^{(n)} g - (C_{XX} + \varepsilon_n I)^{-1} C_{XY} g\|_{\mathcal{H}_X} = O_p(\varepsilon_n^{-1} n^{-1/2}) \quad (n \rightarrow \infty). \quad (8)$$

Since  $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$  for any invertible operators  $A$  and  $B$ , the left hand side is upper bounded by

$$\begin{aligned} & \|(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}(C_{XX} - \widehat{C}_{XX}^{(n)})(C_{XX} + \varepsilon_n I)^{-1}C_{XY}g\|_{\mathcal{H}_X} \\ & \quad + \|(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}(\widehat{C}_{XY}^{(n)} - C_{XY})g\|_{\mathcal{H}_X}. \end{aligned}$$

From  $C_{XY}g = C_{XX}^{\nu+1}\eta$ , we have  $\|(C_{XX} + \varepsilon_n I)^{-1}C_{XY}g\| \leq \|C_{XX}^{\nu}\eta\|_{\mathcal{H}_X}$ . Combination of this fact with  $\|\widehat{C}_{XX}^{(n)} - C_{XX}\| = O_p(n^{-1/2})$  proves that the first term is of the order  $O_p(\varepsilon_n^{-1}n^{-1/2})$ . The second term is of the same order from  $\|\widehat{C}_{XY}^{(n)} - C_{XY}\| = O_p(n^{-1/2})$ , which implies Eq. (8).

Next, we derive the upper bounds

$$\|(C_{XX} + \varepsilon_n I)^{-1}C_{XY}g - E[g(Y)|X = \cdot]\|_{\mathcal{H}_X} = \begin{cases} O(\varepsilon_n^{\nu}), & \text{for } 0 \leq \nu < 1, \\ O(\varepsilon_n), & \text{for } \nu \geq 1. \end{cases} \quad (9)$$

It follows from  $E[g(Y)|X = \cdot] = C_{XX}^{\nu}\eta$  and  $C_{XY}g = C_{XX}^{\nu+1}\eta$  that

$$(C_{XX} + \varepsilon_n I)^{-1}C_{XY}g - E[g(Y)|X = \cdot] = (C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\eta.$$

Let  $C_{XX} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot \rangle$  be the eigendecomposition of  $C_{XX}$  such that  $\lambda_i > 0$  are the eigenvalues and  $\phi_i$  are the orthonormal eigenvectors. The eigenspectrum of the operator  $(C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\eta$  is then given by

$$\frac{\lambda_i^{\nu+1}}{\lambda_i + \varepsilon_n} - \lambda_i^{\nu} = \frac{\lambda_i^{\nu} \varepsilon_n}{\lambda_i + \varepsilon_n} \quad (i = 1, 2, \dots).$$

If  $0 \leq \nu < 1$ , from  $\frac{\lambda^{\nu} \varepsilon_n}{\lambda + \varepsilon_n} = \varepsilon_n^{\nu} \frac{\lambda^{\nu} \varepsilon_n^{1-\nu}}{\lambda + \varepsilon_n} \leq \varepsilon_n^{\nu} \frac{\varepsilon_n^{1-\nu}}{(\lambda + \varepsilon_n)^{1-\nu}}$  and  $|\frac{\varepsilon_n^{1-\nu}}{(\lambda + \varepsilon_n)^{1-\nu}}| \leq 1$  we have

$$\|(C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\eta\| \leq \varepsilon_n^{\nu}.$$

If  $\nu \geq 1$ , then  $\frac{\lambda^{\nu} \varepsilon_n}{\lambda + \varepsilon_n} \leq \varepsilon_n \frac{\lambda^{\nu}}{\lambda + \varepsilon_n} \leq \varepsilon_n \lambda^{\nu-1}$ . It follows

$$\|(C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\eta\| \leq \varepsilon_n \|C_{XX}\|^{\nu-1}.$$

From Eqs. (8) and (9), the proof is completed.  $\square$

## B Proof of Theorem 2

Let  $g_a = \frac{\partial k_{\mathcal{X}}(\cdot, x)}{\partial x^a}$ . Since

$$\begin{aligned} M_{ab}(x) &= \left\langle \langle E[k_{\mathcal{Y}}(*, Y)|X = \cdot], g_a \rangle_{\mathcal{H}_X}, \langle E[k_{\mathcal{Y}}(*, Y)|X = \cdot], g_b \rangle_{\mathcal{H}_X} \right\rangle_{\mathcal{H}_Y} \\ &= \langle E[k_{\mathcal{Y}}(*, Y)|g_a(X)], E[k_{\mathcal{Y}}(*, Y)|g_b(X)] \rangle_{\mathcal{H}_Y} \end{aligned}$$



and

$$\widehat{M}_{n,ab}(x) = \langle \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_a, \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_b \rangle_{\mathcal{H}_Y},$$

we have

$$\begin{aligned} & |\widehat{M}_{n,ab}(x) - M_{ab}(x)| \\ & \leq |\langle \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_a, \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_b - E[k_Y(\cdot, Y)|g_b(X)] \rangle_{\mathcal{H}_Y}| \\ & + |\langle \widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_a - E[k_Y(\cdot, Y)|g_a(X)], E[k_Y(\cdot, Y)|g_b(X)] \rangle_{\mathcal{H}_Y}|. \end{aligned}$$

Noting  $\varepsilon_n \sqrt{n} \rightarrow \infty$  and the expression

$$(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} = (C_{XX} + \varepsilon_n I)^{-1} \{I - (C_{XX} - \widehat{C}_{XX}^{(n)})(C_{XX} + \varepsilon_n I)^{-1}\}^{-1},$$

Lemma 4 in [26] shows that

$$\|C_{XX}(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}\| = O_p(1).$$

From  $g_a = C_{XX}^{\beta+1} \eta$  for some  $\eta \in \mathcal{H}_X$ , we have  $\|\widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g_a\| = O_p(1)$ . For the proof of the first assertion of Theorem 2, it is then sufficient to prove the following theorem.

**Theorem 4.** Assume that  $g \in \mathcal{H}_X$  satisfies  $\mathcal{R}(C_{XX}^{\beta+1})$  for some  $\beta \geq 0$  and that  $E[k_Y(y, Y)|X = \cdot] \in \mathcal{H}_X$  for every  $y \in \mathcal{Y}$ . Then, for  $\varepsilon_n > 0$  with  $\varepsilon_n = n^{-\max\{\frac{1}{3}, \frac{1}{2(\beta+1)}\}}$ , we have

$$\|\widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I_n)^{-1} g - E[k_Y(\cdot, Y)|g(X)]\|_{\mathcal{H}_Y} = O_p\left(n^{-\min\{\frac{1}{3}, \frac{2\beta+1}{4\beta+4}\}}\right)$$

as  $n \rightarrow \infty$ .

*Proof.* It suffices to show

$$\|\widehat{C}_{YX}^{(n)} (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g - C_{YX} (C_{XX} + \varepsilon_n I)^{-1} g\|_{\mathcal{H}_Y}^2 = O_p(\varepsilon_n^{-1/2} n^{-1/2}) \quad (10)$$

and

$$\|C_{YX} (C_{XX} + \varepsilon_n I)^{-1} g - E[k_Y(\cdot, Y)|g(X)]\|_{\mathcal{H}_Y}^2 = O(\varepsilon_n^{\min\{1, (2\beta+1)/2\}}) \quad (11)$$

as  $n \rightarrow \infty$ . In fact, optimizing the rate derives the assertion of the theorem.

Let  $g = C_{XX}^{\beta+1}h$ , where  $h \in \mathcal{H}_X$ . Since  $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$  for any invertible operators  $A$  and  $B$ , the left hand side of Eq. (10) is upper bounded by

$$\begin{aligned} & \|\widehat{C}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}(C_{XX} - \widehat{C}_{XX}^{(n)})(C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\beta+1}h\|_{\mathcal{H}_Y} \\ & + \|(\widehat{C}_{YX}^{(n)} - C_{YX})(C_{XX} + \varepsilon_n I)^{-1}C_{XX}^{\beta+1}h\|_{\mathcal{H}_Y}. \end{aligned}$$

By the decomposition  $\widehat{C}_{YX}^{(n)} = \widehat{C}_{YY}^{(n)1/2}\widehat{W}_{YX}\widehat{C}_{XX}^{(n)1/2}$  with  $\|\widehat{W}_{YX}\| \leq 1$  ([22]), we have  $\|\widehat{C}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}\| = O(\varepsilon_n^{-1/2})$ . It is known that  $\|C_{XX} - \widehat{C}_{XX}^{(n)}\| = O_p(n^{-1/2})$ . From these two fact, we see that the first term is of  $O_p(\varepsilon_n^{-1/2}n^{-1/2})$ . Since the second term is of  $O_p(n^{-1/2})$ , Eq. (10) is obtained.

For Eq. (11), first note that for each  $y$

$$\begin{aligned} E[k_Y(y, Y)|g(X)] &= \langle E[k_Y(y, Y)|X = \cdot], g \rangle = \langle E[k_Y(y, Y)|X = \cdot], C_{XX}^{\beta+1}h \rangle \\ &= \langle C_{XX}E[k_Y(y, Y)|X = \cdot], C_{XX}^{\beta}h \rangle = \langle C_{YX}k_Y(y, \cdot), C_{XX}^{\beta}h \rangle \\ &= \langle k_Y(y, \cdot), C_{YX}C_{XX}^{\beta}h \rangle = (C_{YX}C_{XX}^{\beta}h)(y), \end{aligned}$$

which means  $E[k_Y(\cdot, Y)|g(X)] = C_{YX}C_{XX}^{\beta}h$ . Let  $C_{YX} = C_{YY}^{1/2}W_{YX}C_{XX}^{1/2}$  be the decomposition with  $\|W_{YX}\| \leq 1$ . Then, we have

$$\begin{aligned} & \|C_{YX}(C_{XX} + \varepsilon_n I)^{-1}g - E[k_Y(\cdot, Y)|g(X)]\|_{\mathcal{H}_Y} \\ &= \|C_{YY}^{1/2}W_{YX}\| \|C_{XX}^{\beta+3/2}(C_{XX} + \varepsilon_n I)^{-1}h - C_{XX}^{\beta+1/2}h\|_{\mathcal{H}_Y}. \end{aligned}$$

Let  $\{\phi_i\}$  be the unit eigenvectors of  $C_{XX}$  such that  $C_{XX}f = \sum_i \lambda_i \langle \phi_i, f \rangle \phi_i$ . Then the eigenspectrum of  $C_{XX}^{\beta+3/2}(C_{XX} + \varepsilon_n I)^{-1} - C_{XX}^{\beta+1/2}$  is given by

$$-\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} \quad (i = 1, 2, \dots).$$

If  $0 \leq \beta < 1/2$ , we have  $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} = \frac{\lambda_i^{(2\beta+1)/2}}{(\lambda_i + \varepsilon_n)^{(2\beta+1)/2}} \frac{\varepsilon_n^{(1-2\beta)/2}}{(\lambda_i + \varepsilon_n)^{(1-2\beta)/2}} \varepsilon_n^{(2\beta+1)/2} \leq \varepsilon_n^{(2\beta+1)/2}$ . If  $\beta \geq 1/2$ , then  $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} \leq \lambda_i^{\beta-1/2} \varepsilon_n$ . We have thus Eq. (11), which completes the proof of Theorem 4

□

For the second assertion of Theorem 2, note

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \widehat{M}_n(X_i) - E[M(X)] \right\|_F \\ & \leq \left\| \frac{1}{n} \sum_{i=1}^n \widehat{M}_n(X_i) - \frac{1}{n} \sum_{i=1}^n M(X_i) \right\|_F + \left\| \frac{1}{n} \sum_{i=1}^n M(X_i) - E[M(X)] \right\|_F. \end{aligned}$$

The second term in the right hand side is of  $O_p(n^{-1/2})$  by the central limit theorem. By replacing  $h$  by  $\frac{1}{n} \sum_{i=1}^n h_x^a$  in the proof of Theorem 4, the assertion is obtained as a corollary.

## References

- [1] S. Akaho. A kernel method for canonical correlation analysis. In *Proc. Intern. Meeting on Psychometric Society (IMPS2001)*, 2001.
- [2] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.
- [3] R. Dennis Cook and S. Weisberg. Discussion of Li (1991). *J. Amer. Stat. Assoc.*, 86:328–332, 1991.
- [4] J. Fan and I Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, 1996.
- [5] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *JMLR*, 2:243–264, 2001.
- [6] A. Frank and A. Asuncion. UCI machine learning repository, [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. 2010.
- [7] K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*, 5:73–99, 2004.
- [8] K. Fukumizu, F.R. Bach, and M.I. Jordan. Kernel dimension reduction in regression. *Ann. Stat.*, 37(4):1871–1905, 2009.
- [9] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in NIPS 20*, pages 585–592. 2008.

- [10] A. Gretton, A.J. Smola, O. Bousquet, R. Herbrich, A. Belitski, M.A. Augath, Y. Murayama, J. Pauls, B. Scholkopf, and N.K. Logothetis. Kernel constrained covariance for dependence measurement. In *Proc. AISTATS*, 2005.
- [11] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Stat.*, 29(6):1537–1566, 2001.
- [12] B. Li, H. Zha, and F. Chiaromonte. Contour regression: A general approach to dimension reduction. *Ann. Stat.*, 33(4):1580–1616, 2005.
- [13] K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Stat. Assoc.*, 86:316–342, 1991.
- [14] K.-C. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma. *J. Amer. Stat. Assoc.*, 87:1025–1039, 1992.
- [15] I. Rish, G. Grabarnik, G. Cecchi, F. Pereira, and G.J. Gordon. Closed-form supervised dimensionality reduction with generalized linear models. *Proc. ICML 2008*, pp. 832–839, 2008.
- [16] A.M. Samarov. Exploring regression structure using nonparametric functional estimation. *J. Amer. Stat. Assoc.*, 88(423):836–847, 1993.
- [17] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proc. ICML2009*, pages 961–968. 2009.
- [18] L. Song, A. Smola, K. Borgwardt, and A. Gretton. Colored maximum variance unfolding. In *Advances in NIPS 20*, pages 1385–1392. 2008.
- [19] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.
- [20] M. Wang, F. Sha, and M. Jordan. Unsupervised kernel dimension reduction. *Advances in NIPS 23*, pages 2379–2387. 2010.
- [21] Y. Xia, H. Tong, W.K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *J Royal Stat. Soc., Ser. B*, 64(3):363–410, 2002.
- [22] C.R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.

- [23] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- [24] F. Bauer, S. Pereverzev and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.
- [25] S. Smale, D.. Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, Vol. 19, No. 3. (November 2005), pp. 285-302.
- [26] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.